Recommending Scientific Papers: Investigating the User Curriculum

Jonathas Magalhães
Federal University of Campina Grande
Rua Aprígio Veloso, 882, CEP 58429-900
Campina Grande, Brazil

Cleyton Souza
IFPB – Campus Monteiro
Ac. Rodovia PB - 264, CEP 58500-000
Monteiro, Brazil

Evandro Costa
Federal University of Alagoas
Av. Lourival Melo Mota, S/N, CEP 57072-970
Maceió, Brazil

Joseana Fechine
Federal University of Campina Grande
Rua Aprígio Veloso, 882, CEP 58429-900
Campina Grande, Brazil

Abstract

In this paper, we propose a Personalized Paper Recommender System, a new user-paper based approach that takes into consideration the user academic curriculum vitae. To build the user profiles, we use a Brazilian academic platform called CV-Lattes. Furthermore, we examine some issues related to user profiling, such as (i) we define and compare different strategies to build and represent the user profiles, using terms and using concepts; (ii) we verify how much past information of a user is required to provide good recommendations; (iii) we compare our approaches with the state-of-art in paper recommendation using the CV-Lattes. To validate our strategies, we conduct a user study experiment involving 30 users in the Computer Science domain. Our results show that (i) our approaches outperform the state-of-art in CV-Lattes; (ii) concepts profiles are comparable with the terms profiles; (iii) analyzing the content of the past four years for terms profiles and five years for concepts profiles achieved the best results; and (iv) terms profiles provide better results but they are slower than concepts profiles, thus, if the system needs real time recommendations, concepts profiles are better.

Introduction

In the last decades, there has been a growth in the development of scientific research, consequently increasing the number of published scientific papers (Zhang and Li 2010). Digital Libraries, e.g., ACM (http://portal.acm.org/) and IEEE (http://ieeexplore.ieee.org/), offer to their users a vast collection of scientific papers. Such collection tends to increase its volume due to the periodicity and the emergence of new publishing media. Digital Libraries, mostly, provide search tools to help their users to find relevant content. However, from user's perspective, building the search query using the right terms is not always a simple task. This happens for different reasons, the user may not have experience in the field or the most relevant articles may not appear in the result list because they do not use the same terms of the query (Sugiyama and Kan 2010).

Recommender Systems have been used to mitigate these problems and help users to find relevant content. The research about Paper Recommender Systems can be classified according to the way the recommendation is performed. Basically, there are two main approaches.(i) *Paper-paper* – the recommendation is based on the similarity among papers, analyzing the citations of a given paper or a set of papers. This approach is not personalized, i.e., the recommendation is not different to different users. (ii) *User-paper* – the objective is recommending papers based on the user preferences, through the analysis of the user content. Differently of the *paper-paper* approach, this one provides personalized recommendation.

In this context, in this paper we propose a Personalized Paper Recommender System (PPRS), a new user-paper based approach that takes into consideration the user curriculum. To build the user profiles, we use a Brazilian academic platform called CV-Lattes, where Brazilian researchers maintain data about them, e.g., accepted papers, research projects, areas of interest, etc. Besides presenting the PPRS, we also investigate some issues related to user profiling. First, we define and compare different strategies to build the content-based user profiles with the information extracted from the CV-Lattes: (i) using terms; and (ii) using concepts. Second, we group the user curriculum content by year, and we verify how much past information of a user is required to provide good recommendations.

To validate our strategies, we conduct a user study experiment involving 30 users in the Computer Science domain. Our aim is to answer the following research questions. (i) How many years of the user curriculum are necessary to use in order to provide great recommendations? (ii) Do our approaches outperform the state-of-art in CV-Lattes? (iii) Is there any difference between the concepts profiles and the terms profiles, regarding the results? and (iv) Which method to choose? Respectively, for each question, our results show that (i) it is not necessary verify all content of the user curriculum to provide good recommendations, only the content of the past four years for terms profiles and five years for concepts profiles is enough; (ii) our approaches outperform the state-of-art in CV-Lattes; (iii) concepts profiles are comparable with the terms profiles, even with less information;

and (iv) terms profiles provide better results but they are slower than concepts profiles, thus, if the system needs real time recommendations, concepts profiles are better.

Related Work

Considering the use of concepts to represent user profiles, Chandrasekaran et al. (2008) propose a representation of the user profiles as trees of concepts and an algorithm for computing the similarity between the user profiles and document profiles using a tree-edit distance measure. The user profile is built using his past publications, for each published document in CiteseerX (http://citeseerx.ist.psu.edu), the associated concepts are retrieved and sorted in descending order by their weights. Then, the weights are summed to create a weighted concept vector representing user's interests. The concepts tree is created propagating the weights from a node until the root node. Finally, the utility of a paper to a user is computed by the tree-edit distance, that is calculated by the cost of modifying the document profile to match the user profile. The closer the two profiles, the lower the cost of the required modifications. The work of Kodakateri Pudhiyaveetil et al. (2009) extends (Chandrasekaran et al. 2008) in creating user profiles using the past documents viewed rather than on authored papers, extending the recommendations to CiteSeerX users as well as authors.

Zhang and Li (2010) also use the concept-tree, they present a hybrid recommender system based on collaborative filtering. They compute the similarity between two users using the tree-edit distance. Then, they employ the spreading activation model to search for users those have similar interests with the target user. Finally, the prediction is an aggregation function using the ratings gave by the similar users. The work of Zhang, Wang, and Li (2008) presents a recommender for scientific literatures based on semantic concept similarity computed from the collaborative tags. User profiles and item profiles are represented by semantic concepts. Given a target user, his neighbourhood are selected by collaborative filtering. Then, content-based filtering approach is used to generate recommendation list from the papers these neighbour users tagged.

Middleton, Shadbolt, and De Roure (2004) present an ontological approach for recommender systems. They define a hybrid recommender system, employing both collaborative and content-based recommendation techniques, and they represent user profiles in ontological terms. They use the IBk algorithm to classify the papers in topics. The user profile is computed daily by correlating previously browsed research papers with their classification, ontological relationships between topics of interest are used to infer other topics of interest, an instance of an interest value for a specific class adds 50% of its value to the super-class. Recommendations are formulated from a correlation between the users' current topics of interest and papers classified as belonging to those topics.

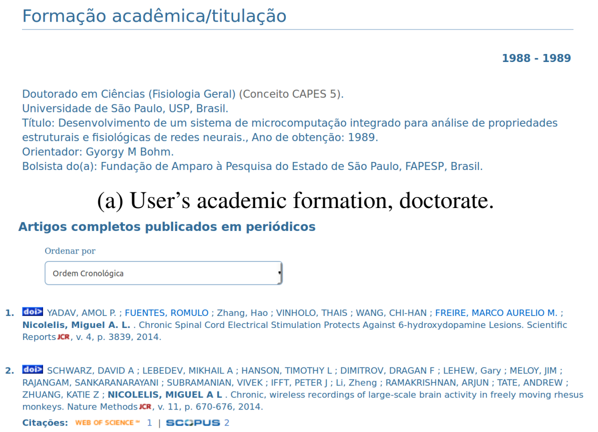
An interesting strategy to compose the user profile is using his past academic production. In this regard, the work of Sugiyama and Kan (2010) considers besides the researcher's past work, they also include the past works' referenced papers as well as papers that cite the work to compose the user profile. They show that using this extra information, it is possible to obtain a more accurate user profile and provide better recommendations, even to the new users. At this stage of our research, we did not perform a comparison with this work, because CV-Lattes does not provide information about the articles cited and referenced. Considering the paper recommendation using the CV-Lattes, we can cite the work of Lopes et al. (2008). They construct the user profile with terms obtained from the "title" and the "keywords" attributes found in the "bibliographic production" and "formation" sections of the user curriculum vitae. Each term has an assigned weight that indicates its importance for the user profile. This weight is obtained by the product of three auxiliary weights: (i) $w_{keyword\_or\_title}$, it considers the term type ("keyword" or "title"), (ii) $w_{language}$, it considers the language of considered term, and (iii) w_{year}, it considers the publication year of the formation or bibliographic production whose term was originated. This work can be considered the state-of-art in CV-Lattes and it is very related to our work, therefore, we compare our approach to this one.

Our work differs from the discussed approaches in the following aspects. Our work utilizes the user curriculum in the recommendation task, we define a mapping function of the curriculum content to create a user profile. Differently of Sugiyama and Kan (2010) and Lopes et al. (2008), we analyze other types of content to build the user profile, e.g., resumé, formation, projects and technical production. From the user viewpoint, our approach can become a very useful service to help in the development of his research. Because our approach converts his effort spent to build his curriculum in to a PPRS. In addition, we perform a comparative analysis of the use of terms and concepts in the representation of user profiles and indexing articles.

The Personalized Paper Recommender System

Definitions – Let $D = \{d_1, ..., d_{|D|}\}$ be the set of all documents in the system. Let $D_u \subseteq D$ be the documents of the user u, where each document $d \in D_u$ has two attributes represented by a tuple $d = (m_d, y_d)$, where m_d represents a textual description of the document and y_d indicates the year that the document was included in the user curriculum. Let $T = \{t_1, ..., t_{|T|}\}$ be the set of terms and $C = \{c_1, ..., c_{|C|}\}$ be the set of concepts, terms and concepts are used to index the documents and to represent user profiles. Each document d is represented by two vectors: (i) using terms $\vec{dt} = (w_{d,1}, ..., w_{d,|T|})$, where the weight $w_{d,t}$ represents the importance of the term t to the document d; and (ii) using concepts $\vec{dc} = (w_{d,1}, ..., w_{d,|C|})$, where the weight $w_{d,c}$ represents the importance of the concept c to the document d. The same way as documents, the users are represented by terms and concepts, so the user u has two profiles: (i) a terms profile denoted by the vector $\vec{pt}_u = (w_{u,1}, ..., w_{u,|T|})$, where $w_{u,t}$ represents the importance of the term t to the user u; and (ii) a concepts profile denoted by $\vec{pc}_u = (w_{u,1}, ..., w_{u,|C|})$, where $w_{u,c}$ represents the weight of the concept c to the user u. In the rest of this section, when we present the equation sim, we are referring to the cosine similarity.

The CV-Lattes – The CV-Lattes (http://lattes.cnpq.br/) was developed by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (Brazilian National Council for Scientific and Technological Development - CNPq – http://www.cnpq.br/) and it is available in Portuguese and English. Its objective is standardize the information about, but not restricted, the Brazilian scientific community. This information is used by Brazilian government agencies to evaluate researches, projects, graduate and postgraduate programs, among others. Therefore, the platform attracts researchers of different levels from undergraduate students to reputed researches worldwide. According to the site PainelLattes (http://estatico.cnpq.br/painelLattes/), on Feb 19th 2015, the CV-Lattes reached the mark of 3,098,215 registered curricula. This set includes researches, students and professionals from different fields of knowledge, where 1,199,734 are from students. Figure 1 presents two parts of a user CV-Lattes (goo.gl/bj4XaJ), the user's academic formation and bibliographical production. In the following, we present some examples of sections available for users to register in their online curriculum. **Resumé** – a short description of the user biography. **Academic Formation** – data from his academic degree, graduation, masters and doctorate. It consists of: type of academic degree, institution, title of the thesis, the years that he began and ended the thesis, and advisors and co-advisors. Figure 1a presents an example of academic formation in CV-Lattes. **Projects** – data about the research projects which the user participates or participated. For each project the user includes the following data: title, years of project duration, description, members, and keywords. **Bibliographical Production** – every type of published material: e.g., papers in conferences, papers in journals, books, book chapters.



Formação acadêmica/titulação

1988 - 1989

Doutorado em Ciências (Fisiologia Geral) (Conceito CAPES 5).
Universidade de São Paulo, USP, Brasil.
Título: Desenvolvimento de um sistema de microcomputação integrado para análise de propriedades estruturais e fisiológicas de redes neurais., Ano de obtenção: 1989.
Orientador: Gyorgy M Bohm.
Bolsista do(a): Fundação de Amparo à Pesquisa do Estado de São Paulo, FAPESP, Brasil.

(a) User's academic formation, doctorate.

Artigos completos publicados em periódicos

Ordenar por

Ordem Cronológica

1. doi> YADAV, AMOL P. ; FUENTES, ROMULO ; Zhang, Hao ; VINHOLO, THAIS ; WANG, CHI-HAN ; FREIRE, MARCO AURELIO M. ; **Nicolelis, Miguel A. L.** . Chronic Spinal Cord Electrical Stimulation Protects Against 6-hydroxydopamine Lesions. Scientific Reports JCR, v. 4, p. 3839, 2014.

2. doi> SCHWARZ, DAVID A ; LEBEDEV, MIKHAIL A ; HANSON, TIMOTHY L ; DIMITROV, DRAGAN F ; LEHEW, Gary ; MELOY, JIM ; RAJANGAM, SANKARANARAYANI ; SUBRAMANIAN, VIVEK ; IFFT, PETER J ; Li, Zheng ; RAMAKRISHNAN, ARJUN ; TATE, ANDREW ; ZHUANG, KATIE Z ; **NICOLELIS, MIGUEL A L** . Chronic, wireless recordings of large-scale brain activity in freely moving rhesus monkeys. Nature Methods JCR, v. 11, p. 670-676, 2014.

Citações: WEB OF SCIENCE 1 | SCOPUS 2

(b) User's bibliographical production, papers published in journals.

Figure 1: Two parts of a user CV-Lattes.

Crawling the CV-Lattes – The user's data is crawled from CV-Lattes and it is saved in the system to create the set D_u of user documents, then, this set is used to build the user profile. The CV-Lattes platform do not provide an API for developers, however the tool scriptLattes (Mena-Chalco and Junior 2009) permits to automatically get information about a user. Thus, using the scriptLattes, we developed a crawler that takes a user's CV-Lattes id, analyzes the user's page and returns the user's data. Table 1 presents the mapping between the data crawled from CV-Lattes to the model representation. Each peace of the user's CV-Lattes is mapped to a document $d \in D_u$, in Table 1 we present the maximum number of documents that can be generated using each type of data. For example, the user has at most three formation documents: graduate, masters and Phd.

Table 1: Mapping the data crawled from CV-Lattes to the model representation.

CV-Lattes Data	**Document attributes**		**Max Number**
	Document Description (m_d)	**Document Year (y_d)**	
Resumé	description	year now	1
Formation	title	year	3
Projects	title + description	conclusion year	n
Technical Production	title + description + keywords	year	n
Bibliographical Production	title + description + keywords	year	n

Pre-Processing Data – This process consists of a Natural Language Processing (NLP) pipeline. For this, we use the tool NLTK (Natural Language Toolkit) (http://nltk.org/) which is used for NLP tasks, e.g., classification, tokenization, stemming, etc. The input is the crawled raw data and the output is the processed data. In the following, we explain the details about the steps of this process. (i) **Translation** – the textual data from user curriculum is translated from Portuguese to English, this was necessary because the papers for recommendation are written in English. In this step, we used the tool Google Translate (goo.gl/SKZX7u). (ii) **Normalization** – special characters are removed (eg.: digits and punctuation) and the text is placed in lowercase. (iii) **Tokenization** – unigrams from text are removed and a list of tokens is generated. (iv) **Stop-words removal** – the *stop-words* are removed. In this step, we use the stop-words available in the NLTK. (v) **Stemming** – application of Lancaster Stemming (Paice 1990).

Building the Knowledge Base – We use a knowledge base to build the user concepts profile and index the papers by concepts. We use an ontology-based approach to represent the domain and we follow the approach proposed by Loh et al. (2006). The set $C = \{c_1, ..., c_{|C|}\}$ represents the concepts associated to the domain, each concept $c \in C$ is a node in the ontology. Each concept $c \in C$ is represented by a term vector, $\vec{c} = (w_{c,1}, ..., w_{c,|T|})$, and, the weight $w_{c,t}$ means how much the term t is related to the concept c. Each concept $c \in C$ has a training set defined by $D_c \subseteq D$ and m_{D_c} represents the concatenation of all descriptions of the documents $d \in D_c$. The weight $w_{c,t}$ is calculated statistically by the TF-IDF scheme, so $w_{c,t} = \text{TF}(t, m_{D_c}) * \text{IDF}(t, \{m_{D_{c_1}}, ..., m_{D_{c_{|C|}}}\})$.

The Terms Profile $\vec{pt}_u$ – The terms profile $\vec{pt}_u$ of the user u is built using the user set D_u crawled from CV-Lattes. The weight $w_{u,t}$ of $\vec{pt}_u$ is defined by:

$$w_{u,t} = \sum_{d \in D_u} \text{TF}(t, m_d) * \text{temp}(y_d), \qquad (1)$$

where the function $\text{temp}(y_d)$ makes a temporal calibration, its objective is giving more importance to recent documents than older ones. The function $\text{temp}(y_d)$ is defined by:

$$\text{temp}(y_d) = -\frac{y_{now} - y_d}{v} + 1, \qquad (2)$$

where $v \in \mathbb{N}^*$ represents the years interval that a content is considered in the user profile, y_{now} is the present year and the y_d is the year which the document was released.

The Concepts Profile $\vec{pc}_u$ – The concepts profile $\vec{pc}_u$ of the user u is defined combining the terms profile $\vec{pt}_u$ with the concepts vectors of the knowledge base. Thus, the weight $w_{u,c}$ of $\vec{pc}_u$ means how much the user u is related to the concept c. It is computed by the similarity between the concept vector $\vec{c}$ and the user terms profile $\vec{pt}_u$, i.e., $w_{u,c} = sim(\vec{pt}_u, \vec{c})$.

Indexing Papers – The available papers for recommendation are indexed by terms and concepts, let the $D^{rec} \subseteq D$ be the set of available papers for recommendation, so each paper $d \in D^{rec}$ is represented by two vectors: (i) **Using terms** $\vec{dt}$, where the weight $w_{d,t}$ of $\vec{dt}$ is given by TF-IDF scheme, so: $w_{d,t} = \text{TF-IDF}(t, d, D^{rec})$. (ii) **Using concepts** $\vec{dc}$, where the weight $w_{d,c}$ of $\vec{dc}$ is computed by the similarity between the terms profile and the concepts profile $\vec{c}$, thus: $w_{d,c} = sim(\vec{dt}, \vec{c})$.

Recommending Papers – The recommendation for a user u is performed analyzing the available papers D^{rec}, the recommendation is given by:

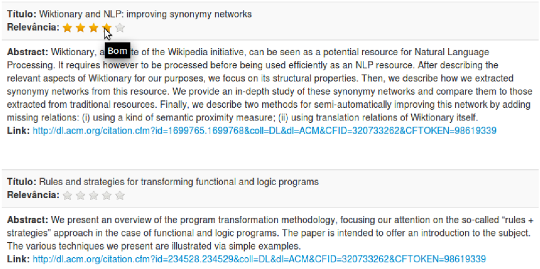
$$D^{rec}_u = \underset{d \in D^{rec} \setminus D_u}{\overset{n}{\text{argmax}}} \; util(u, d), \qquad (3)$$

where the function argmax returns the n more relevant papers d to the user u. The function $util(u, d)$ returns the utility of a document d to the user u. In the following, we define two approaches to compute this utility. **Similarity between terms profiles**: $util(u, d) = sim(\vec{pt}_u, \vec{dt})$. **Similarity between concepts profiles**: $util(u, d) = sim(\vec{pc}_u, \vec{dc})$.

Methodology

Setting Up the System – To validate our proposal, we conduct a user study experiment to create a benchmark. We develop a Paper Recommender System in the Computer Science domain. The system enables users to register in the system and inform their information, e.g., name, institution, CV-Lattes id, email and academic degree. The CV-Lattes id consists of the link to the user's CV-Lattes. The users also mark their preferences in a predefined set of 50 papers, our main objective is to built a ground truth dataset, benchmark, to compare the discussed approaches. Figure 2 presents part of papers display, for each paper the user could see title, abstract and an external link to the paper. Then, users rated the papers using a Likert scale, using 1 to 5 stars grade. We present the papers following these steps: (i) we create ten groups of papers labeled from 0 to 9 with five papers each; (ii) the first group displayed to a user was that labeled with the last digit of user's id, and; (iii) we display the other groups of papers in a circular order of the label sequence. For example, the group of papers to the user with id 14 were displayed in the following order: 4-5-6-7-8-9-0-1-2-3. This procedure was adopted to minimize the effect of the display order of the papers in the results. The users rated the papers using stars, with the following meaning. One star: *Inadequado* (Inadequate); two stars: *Ruim* (Bad); three stars: *Médio* (Average); four stars: *Bom* (Good) and five stars: *Excelente* (Excellent).

Figure 2: Part of the display screen of papers.



Título: Wiktionary and NLP: improving synonymy networks
Relevância: ★★★★☆

Abstract: Wiktionary, a te of the Wikipedia initiative, can be seen as a potential resource for Natural Language Processing. It requires however to be processed before being used efficiently as an NLP resource. After describing the relevant aspects of Wiktionary for our purposes, we focus on its structural properties. Then, we describe how we extracted synonymy networks from this resource. We provide an in-depth study of these synonymy networks and compare them to those extracted from traditional resources. Finally, we describe two methods for semi-automatically improving this network by adding missing relations: (i) using a kind of semantic proximity measure; (ii) using translation relations of Wiktionary itself.
Link: http://dl.acm.org/citation.cfm?id=1699765.1699768&coll=DL&dl=ACM&CFID=320733262&CFTOKEN=98619339

Título: Rules and strategies for transforming functional and logic programs
Relevância: ☆☆☆☆☆

Abstract: We present an overview of the program transformation methodology, focusing our attention on the so-called "rules + strategies" approach in the case of functional and logic programs. The paper is intended to offer an introduction to the subject. The various techniques we present are illustrated via simple examples.
Link: http://dl.acm.org/citation.cfm?id=234528.234529&coll=DL&dl=ACM&CFID=320733262&CFTOKEN=98619339

When a user passed the mouse over the stars he was informed about the meaning of the number of stars according to our scale. For example, in Figure 2, the user is marking a paper with four stars and she reads *Bom* that means "Good".

Setting the Knowledge Base – The knowledge base was constructed using 19 concepts based on Mendeley Computer and Information Science sub-disciplines (goo.gl/KqDtbI): Algorithms and Computational Theory, Artificial Intelligence, Computer Architecture, Computer Security, Data Communication and Networks, Database Systems, Design Automation, Electronic Commerce, Graphics, Human-Computer Interaction, Information Retrieval, Information Science, Information Storage, Multimedia Systems and Applications, Operating Systems, Programming Languages, Real-Time Systems, Software Engineering, Systems and Control Theory. For each concept, we define a data set D_c containing 1000 papers using the Mendeley API (http://apidocs.mendeley.com/), using the concept as query. First, we concatenated and pre-processed the title and abstract of each paper, then, we compute the concepts vectors.

Setting the Papers and User Profiles – The set of the available papers to the users consists of 50 Computer Science papers, 25 papers related to Artificial Intelligence and 25 papers related to Software Engineering. They were proposed by two specialists, Phd Evandro Costa (http://goo.gl/98ljK6) (Artificial Intelligence) and Phd Baldoino Fonseca (http://goo.gl/oCs9k8) (Software Engineering). The papers are available in (http://goo.gl/IHHvqI), including their specification in relation to the subareas of Artificial Intelligence and Software Engineering. We defined the subareas according to ACM Taxonomy from 1998 (goo.gl/V4lf7T). For each paper we pre-process the concatenation of its title and abstract, then we indexed the papers. We also ask to the specialists to select papers from well quoted conferences to isolate the quality factor of the paper in the experiment. Thus, we expect to prevent that a paper that a user considers relevant receives a low rating because its quality. We use the user's CV-Lattes id to get user data from Lattes and compose his set of documents D_u. Then, we pre-process the user data, then, we build the user profiles using the documents D_u.

Experimental Factors – In our experiment we use two factors. **Recommendation method** – levels: (i) recommendation using the user terms profile (TP); (ii) recommendation using the concepts profile (CP); and (iii) the recommender system proposed by Lopes et al. (Lopes et al. 2008) ($Lopes$). **Years considered in user profiles (v)** – this factor is used in Equation 2 and defines if a content will be used or not in the user profile. It has the following levels: (i) 1 year; (ii) 2 years; (iii) 3 years; (iv) 4 years; (v) 5 years; and (vi) all years, i.e., all content in user profile. In the text, we cite a profile using the abbreviation, $type + v$, e.g., $TP5$ means the terms profile with $v = 5$.

Metrics – We compare the methods using the Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen 2002). The NDCG associates the position of the item in the rank and its relevance. The better is the position of the most relevant items, the higher is the value of NDCG, being the optimal value 1. We are interested in few items in the recommendation list, so we use NDCG@N (N = 5, 10) to evaluate just the top-N papers in the recommendation list. So, we present the NDCG average over all users. We also evaluate the profile length, i.e., the average number of terms in the user profile, the intuition is to verify how simpler is the profile.

Questions – We attempt to answer the following questions. First, we want to verify how much data from user curriculum is necessary to provide great recommendations, so our first question is Q_1 – How many years (v) of the user curriculum are necessary to use in order to provide great recommendations (Metrics: NDCG@5 and NDCG@10)? Second, we compare our two strategies to represent the user profiles, thus Q_2 – Are the proposed methods (TP and CP) better than $Lopes$ (Metrics: NDCG@5 and NDCG@10)? Then, we want to compare our strategies with the state-of-art in CV-Lattes, so Q_3 – Is there difference between the concepts profile CP and the terms profile TP (Metrics: NDCG@5 and NDCG@10)? Finally, we want to verify what is the best method, thus Q_4 – Which method to choose (TP, CP or $Lopes$) (Metrics: NDCG@5 and Length)?

Results and Discussion

Results – We disclosed the system through Computer Science email lists of four Brazilian universities, a total of 73 users attended to the system but only 30 rated all papers, thus, we only use these 30 users in the validation. Grouping by institution, we had 14 users from Federal University of Alagoas, 9 users from Federal University of Campina Grande, 3 users from Federal University of Pernambuco and 4 users from others. Considering the user academic level, we have 9 undergraduate students, 2 graduated, 8 Msc students, 2 Msc, 8 Phd students and 1 Phd. In the following, we present the average number of each information type in the users curriculum, resumé = 0.9393, formation = 0.9, projects = 2.833, technical production = 6.167 and bibliographical production = 8.633.

Table 2 presents the NDCG@5, NDCG@10 average for all recommendation methods. It also presents the length average for the terms profiles generated. Table 3 presents the statistical tests comparing the strategies. We choose the statistical test according to data distribution, if it is normal then Student's t test, otherwise Wilcoxon's test.

Table 2: The NDCG@5, NDCG@10 and length means of the methods of the generated profiles. We execute the Shapiro-Wilk test to verify the data normality. The symbol (*) indicates that the data is not normally distributed, i.e., $p\text{-}value < 0.05$.

Metric	**NDCG@5**		**NDCG@10**		**Length**
Years	**Type**				
(v)	**TP**	**CP**	**TP**	**CP**	**TP**
1	0.3701*	0.3239*	0.3722*	0.3448*	29.63*
2	0.4047	0.3503*	0.4259	0.3793	45.17
3	0.4698	0.4008	0.4787	0.4398	57.13
4	**0.4750**	0.4206	**0.4806**	0.4428	71.1
5	0.4613	**0.4435**	0.4608	**0.4569**	76.90
All	0.4560	0.4234	0.4529	0.4530	82.63
Lopes	0.3250*		0.3489*		46.27*

Table 3: Results of hypothesis tests performed to compare the strategies. Both tests are performed with parameters $\alpha = 0.05$, alternative = "greater", paired = TRUE ("$\gg$" and "$>$" denote significance levels of $p\text{-}value < 0.01$ and $p\text{-}value < 0.05$, respectively).

Metric	**Student's t test**	**Wilcoxon's test**
NDCG@5	$TP4 > TP2$	$TP4 > (CP1, CP2)$
	$TP4 > CP3$	$TP4 \gg Lopes)$
	$CP5 > CP3$	$CP5 > (CP1, Lopes)$
NDCG@10	$TP4 \gg TP5$	$TP4 \gg (CP1, Lopes)$
	$TP4 > CP2$	$TP4 > TP1$
	$TP4 > CPall$	$CP5 \gg CP1$
	$CP5 > CP2$	$CP5 > (TP1, Lopes)$

Discussion – In the following, we answer our research questions.

Q_1 **– How many years (v) of the user curriculum are necessary to use in order to provide great recommendations?** Analyzing Tables 2 and 3 we verify that, for both methods, it was not necessary to verify all data of the user curriculum, e.g., $TP4 \gg TP5$ and $TP4 > CPall$. The main reason is that users have marked the items according to their current preferences. Thus, the old contents should be cut or have their weights decreased by the forgetting factor. We can conclude that the answer for Q_1 is: *It depends of the profile type, fours years for TP and five years for CP.* This an interesting result and supports the results of Cremonesi, Milano, and Turrin (2012). They investigate how many ratings should be collected from a new user before providing recommendations They observe that profile lengths longer than 10 ratings do not increase user perceived relevance in the recommendations.

Q_2 **– Are the proposed methods (CP and TP) better than $Lopes$?** According to the statistical tests, for both metrics, the proposed methods $CP5$ and $TP4$ are significantly better than method $Lopes$. This occurs mainly because we analyze more information of the CV-Lattes, e.g., resumé, formation, projects and technical production, thus building a more accurate profile. So, we can answer the question Q_2: *Yes, our approaches ($TP4$ and $CP5$) achieved better performance than $Lopes$.*

Q_3 **– Is there difference between the concepts profile CP and the terms profile TP?** Comparing the terms profile $TP4$ to the concepts profile $CP5$ (Table 2), we verify that the $TP4$ achieved a better mean of NDCG@5 and NDCG@10. We perform a Paired Student's t test ($\alpha = 0.05$, $H_a : CP5 \neq TP4$) to verify if the difference between the means is significant. We obtain a $p\text{-}value = 0.3675$ for NDCG@5 and $p\text{-}value = 0.4797$ for NDCG@10. Thus, $TP4$ is not statistically better than CP. We credit this fact mainly to the quality of the knowledge base. Thus, we can answer the Question Q_3: *Yes, the concepts profile $CP5$ obtained a recommendation quality statistically comparable with the method $TP4$.*

Q_4 **– Which method to choose (TP, CP or $Lopes$)?** To answer this question we consider other aspects besides the quality of the recommendation, in the following we present advantages and disadvantages of the methods TP and CP. TP – This method has the advantage of being easily adaptable to other domains because it does not require knowledge base, however depending on the quantity of terms used (Table 2), in the recommendation process, the similarity calculation consumes more time. CP – This method involves a knowledge engineering to define the concepts and the training set to build the knowledge base. This fact implies a more time-consuming and laborious construction profile, besides having a difficult adaptation to other domains. However, after this procedure performed and proven its effectiveness, this profile has some advantages, such as: (i) it computes the recommendation faster than TP, because the vector of weights is smaller; and (ii) it is simpler, i.e., better for the user to inspect it.

After these analysis, we can answer the question Q_4: *It depends on the context, because there is a trade-off between the techniques. If the system needs an online recommendation with reasonable quality, the CP profiles are the best choice. On the other hand, if the systems can compute the recommendations offline, and the time consuming is not a problem, the TP is better.* However, we emphasize that the profile CP is composed of only 19 concepts, being necessary to investigate other ways of obtaining knowledge domain and the use of a higher number of concepts.

Conclusion and Future Work

In this paper, we presented and evaluated our approach to a paper Recommender System that considers the user curriculum crawled from the CV-Lattes. Our main contributions are: (i) in regard to the recommendation quality we obtained better results in comparison to state-of-art in CV-Lattes; (ii) we showed that the concepts profiles can be statistically comparable with the terms profiles; and (iii) we built a knowledge base in Computer Science that can be used by other works. For future work we are planning: i) to extend our approach to researches that do not have a CV-Lattes page; (ii) to incorporate and integrate data from other systems into the user profile, e.g., Mendeley, LinkedIn; iii) to improve the recommender model with other attributes of the papers, e.g., cited and referred papers; and (iv) to develop an online tool to recommend papers to the users.

References

Chandrasekaran, K.; Gauch, S.; Lakkaraju, P.; and Luong, H. P. 2008. Concept-based document recommendations for citeseer authors. In *Proceedings of the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, AH '08, 83–92. Berlin, Heidelberg: Springer-Verlag.

Cremonesi, P.; Milano, P.; and Turrin, R. 2012. User effort vs. accuracy in rating-based elicitation. In *Proceedings of the 6th ACM conference on Recommender systems - RecSys '12*, 27–34.

Järvelin, K., and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.* 20(4):422–446.

Kodakateri Pudhiyaveetil, A.; Gauch, S.; Luong, H.; and Eno, J. 2009. Conceptual recommender system for citeseerx. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, 241–244. New York, NY, USA: ACM.

Loh, S.; Lichtnow, D.; Borges, T.; Piltcher, G.; and Freitas, M. 2006. Constructing domain ontologies for indexing texts and creating users' profiles. In *Work. on Ontologies and Metamodeling in Software and Data Engineering, Brazilian Symp. on Databases, UFSC, Florianópolis*, number 2003, 72–82.

Lopes, G. R.; Souto, M. A. M.; Wives, L. K.; and de Oliveira, J. P. M. 2008. A personalized recommender system for digital libraries. In *Proceedings of the 14th Brazilian Symposium on Multimedia and the Web*, WebMedia '08, 59–66. New York, NY, USA: ACM.

Mena-Chalco, J., and Junior, R. 2009. scriptLattes: an open-source knowledge extraction system from the Lattes platform. *Journal of the Brazilian Computer Society* 15(4):31–39.

Middleton, S. E.; Shadbolt, N. R.; and De Roure, D. C. 2004. Ontological user profiling in recommender systems. *ACM Trans. Inf. Syst.* 22(1):54–88.

Paice, C. D. 1990. Another stemmer. *SIGIR Forum* 24(3):56–61.

Sugiyama, K., and Kan, M.-Y. 2010. Scholarly paper recommendation via user's recent research interests. In *Proceedings of the 10th annual joint conference on Digital libraries*, JCDL '10, 29–38. New York, NY, USA: ACM.

Zhang, Z., and Li, L. 2010. A research paper recommender system based on spreading activation model. In *Information Science and Engineering (ICISE), 2010 2nd International Conference on*, 928–931. IEEE.

Zhang, M.; Wang, W.; and Li, X. 2008. A paper recommender for scientific literatures based on semantic concept similarity. In Buchanan, G.; Masoodian, M.; and Cunningham, S., eds., *Digital Libraries: Universal and Ubiquitous Access to Information*, volume 5362 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 359–362.